USING MARKOV CHAIN ON ONLINE LEARNING HISTORY DATA TO DEVELOP LEARNER MODEL FOR MEASURING STRENGTH OF LEARNING HABITS

Tuan M. Tran[1] and Shinobu Hasegawa[2]
[1]*School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Nomi City 923-1292, Japan*
[2]*Center for Innovation Distance Education and Research, Japan Advanced Institute of Science and Technology (JAIST), Nomi City 923-1292, Japan*

ABSTRACT

A learner model reflects learning patterns and characteristics of a learner. A learner model with learning history and its effectiveness plays a significant role in supporting a learner's understanding of their strengths and weaknesses of their way of learning in order to make proper adjustments for improvement. Nowadays, learners have been engaging in online learning frequently and intensely, leaving behind tremendous learning history data that contain informative insights about the learners' learning patterns. This paper proposed a method for developing learner models by applying the Markov chain to learning history data. Our method transforms individual learners' resource use data in a learning course into a large amount of resource use sequences, then develops a Markov learner model, and generates the resource use steady state for each learner. The resource use density, the resource steady state, and the assessment scores of individual learners tell their learning patterns and the effectiveness of the learning patterns. From the Markov learner model, we generate a learner profile for describing learning patterns and an index for measuring the strength of learning habits of the learner. We verified our method by applying it to each course in the OULAD dataset to predict the learning performance using the index. The preliminary results gain up to 97% accuracy on the pass/fail prediction problem.

KEYWORDS

Clickstream Data, Online Learning Environment, Markov Chain, Learner Model, SRL Profile, SRL Index

1. INTRODUCTION

Online learning has a weighty contribution to the learning process. Nowadays, it is almost requisite that education institutions deliver learning materials online for students to self-study besides classroom times. Students' online learning activities produce undoubtedly gigantic data that somehow reveal and describe learning patterns by which the students use online resources for learning. Learner models are means to describe learning patterns. The characteristics of online learning bring an opportunity to generate learner models that can reflect learners' ways of learning adequately based on learning data about the learners themselves, therefore, helping learners know their learning patterns and gradually become aware of their strengths and weaknesses. Certainly, learning habits have a significant impact on the learning process. Good learning habits enhance the learning process. Bad ones prevent it. Learning patterns are made up of activity repetition; hence, they can be used to derive learning habits. This research aims to model learning habits and their effectiveness on learning performance. The contribution of this paper is a learner model that is generated from learning history data, able to reflect learning habits, and measurable. This contribution is accomplished by three main elements: the *Markov learner model* for students' online learning patterns, the *self-regulated learning (SRL) profile* to visualize the model, and *the SRL index* as a scalar measurement for the profile. Learners usually self-regulate their learning behavior consciously or unconsciously when engaging resources in OLEs (Bernacki et al., 2011). Therefore, the SRL prefix is used for the profile and the index.

2. METHODOLOGY

We will briefly review current research on the most popular type of online learning data known as clickstream, then describe the Markov chain and how to apply it to the learner model.

Online learning clickstream data mainly contain mouse clicks on learning resources, who click the resources and timestamps when they happen. A number of current research focus on clickstream data not only to predict students' performance in online learning environments (OLEs) but also to establish learning behavior models. Specifically, there have been applications of the Markov chain to modeling learning patterns and hidden learning stages of learners (Geigle & Zhai, 2017; Qiao et al., 2021; Waheed et al., 2020).

A Markov chain (Bertsekas & Tsitsiklis, 2000) is a sequence of states appearing through time with a special property in which the value of a state only depends on a state right before it and is independent of other previous states. It is called the Markov property. Learning in an OLE manifests as a sequence of access to resources provided by the environment. Learners access resources by various modes but the most recognizable type of access is mouse clicking on the resources and their inner content. OLEs usually present several resources at once, and a learner can freely use the resources in various orders. From these observations, a learning process in OLEs can be seen as a stochastic process with resources as its states.

In a Markov model, the probability distribution over states approaches stability as state transition happens a large number of times. This condition is known as the steady state where the probability distribution over states is independent of the likelihood each of the states initiates a sequence; therefore, a steady state is like a habit of that which a Markov model represents (Gilbert Strang, 2016).

The current existing studies produce learner models from collective data of groups of learners rather than from data about individual learners. In addition, the current models seemed not to reach a steady state to tell the enduring characteristics of learning patterns. We think that it is beneficial for individual learners to have their learner model reflect their very own learning patterns. Thus, our research strives for such a learner model.

2.1 Problem Statement

To generate a Markov model for a learner from their learning history, we state the problem as this. From a learner's record of resource use in a course in an OLE, what are the learner's habits of using resources?

In online learning, habits are patterns of resource access. These habits tend to maintain among courses with a similar design in terms of resource types. Therefore, habits that support the learner's success in one course can enable success in other courses of similar design. And habits that hinder learning progress in a certain course might prevent the learner from learning effectively in the others. In the following, we will describe how the Markov model helps analyze records of learning resources and develop the learner model to represent learning habits.

2.2 Arrangement of a Resource Use History to Form Resource Sequences

There are several OLEs being used, different from each other in the kind of data they are storing, but they share similar arrangement and information about resource use records, namely, what resources learners access or click in a certain time unit. Following the shared structure, we can read from a resource use history of learners about their learning pattern comprising of what resources have been used, sequences of resource access, and resource use before and after a certain milestone such as an exam or a lecture.

We developed Markov models from Online University Learning Analytics Dataset, also known as OULAD (Kuzilek et al., 2017). The time step unit in OULAD is a day. A learner often accesses plenty of resources in a day in an unknown order. So, we do not know resource use sequences in a day; however, it is possible to obtain sequences of resources a learner accesses from day to day successively. As shown in Figure 1, the resource use history data, when rearranged (top) and traversed (down) in ascending order of day and show a clearer sequence pattern. A resource use record of one learner, when being observed in this manner, generates a large number of sequences, sufficient to produce a Markov model relevant to the learners' patterns of resource use.

2.3 Markov Learner Model, SRL Profile, and SRL Index

We used the markovchain package (Spedicato, 2017) for R to build Markov learner models from OULAD. To describe and rank the learner model, we generated from the model two units named *SRL profile* and *SRL index*, respectively.

The *SRL profile* shows the frequency and density of resource use in a course, and the effectiveness of resource use in such a pattern. Frequency refers to a probability distribution over resource use, and density the number of access to resources.

The *SRL index* indicates how effective a resource use pattern is in corresponding to learning performance and is expressed in the form *SRL index=* $c\_score \sum_{i=1}^{n} r_i^{ss_i}$ where *c_score* is an accumulative assessment score a learner has earned in the course, r_i the density of resource *i* in the course measured by the number of resource access, ss_i the steady state probability of resource r_i , and *n* the number of resources available in the course. The term $r_i^{ss_i}$ illustrates a steady pattern of use of resource *i*. $\sum_{i=0}^{n} r_i^{ss_i}$ is a summary of resource use pattern of a learner in a course. An effective learning pattern is likely to result in a high cognitive performance, expressed in the term $c\_score \sum_{i=0}^{n} r_i^{ss_i}$. To retain data normality, log transformation is applied to the SRL index as *Log SRL index=* $log_2 (c\_score \sum_{i=1}^{n} r_i^{ss_i})$.

Figure 2 shows an SRL profile example. All four graphs comprise a big picture of learning patterns and SRL abilities of all learners grouped by their final results in a course. There are four final results in this case: *Distinction learners* who have passed the course with the highest scores, *Pass learners* who passed the course, *Fail learners* who have completed but not passed the course, and *Withdrawn learners* who have not completed the course. The big grey circles indicate the current states of the learner who owns this SRL profile. The upper-left boxplot (a) tells that the learner is in the Distinction group with an SRL index of about 11 points. The upper-right scatterplot (b) shows that the learner's cognitive score of more than 90 points in relation to his or her SRL index of about 11 points. The lower-left (c) and lower-right (d) scatterplots demonstrate the number of accesses to each resource and its steady state in the course.

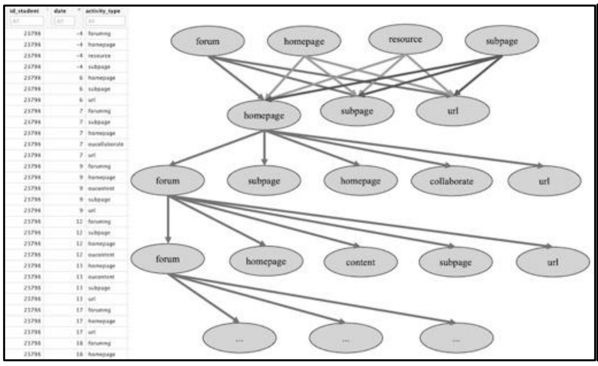

Figure 1. Formation of resource use sequences

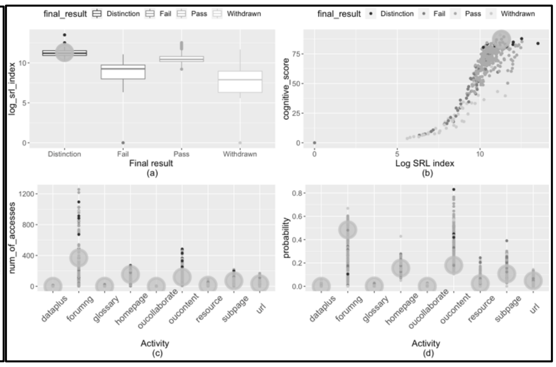

Figure 2. An SRL Profile

3. PRELIMINARY RESULTS

We justified the proposed learner model by developing learning performance prediction models from the SRL index and assessment scores, then compared the prediction accuracies with current state-of-the-art prediction models: *(i)* Hao et al. (2022) analyzing the sum of clicks on resources and average assessments scores to predict pass/fail final results with 93% accuracy and 0.88 F1-score, and *(ii)* Qiu et al. (2022) classifying resource types used by each student in a course to predict pass/fail final results with 97% accuracy and 0.98 F1-score, both on OULAD.

We generated SRL profiles for individual learners from their resource access throughout a whole course, computed the SRL indices, and used the SRL indices and accumulative assessment scores as features to build the machine learning models by the Support vector machine with linear kernel algorithm to predict the final results of pass or fail. Two approaches for splitting training and testing datasets were carried out. The first approach used one course offering and then predict the learning performance of students in other course offerings. The second approach was to split each course offering by a 0.5 training: 0.5 testing data ratio. Table 1 presents the mean, maximum, and minimum values with standard deviation (SD) of accuracy and F1-score of the prediction models developed in two approaches. Both approaches yielded high accuracies with the maximums equivalent to the state-of-the-art.

Table 1

Approach	Accuracy			F1-Score		
	Mean ± SD	Max	Min	Mean ± SD	Max	Min
1st	0.92 ± 0.02	0.94	0.87	0.90 ± 0.03	0.96	0.87
2nd	0.93 ± 0.02	0.97	0.91	0.93± 0.03	0.97	0.89

4. CONCLUSION

The proposed Markov learner model transforms learning history data into a descriptive and measurable learner profile that unifies learning activities, learning patterns, and their effectiveness. Our proposed method is not limited to a specific case like the OULAD, but can be generalized and applicable to almost any OLEs.

Our work has limitations as these. The first limitation relates to the verification of the model. It is important to apply the method to other open datasets and to specific learning history data at educational institutions to validate the method. Second limitation is of the SRL profile. The SRL profile intends to give learners a detailed illustration of their SRL patterns. Looking at their SRL profiles, learners are expected to recognize their SRL habits and be able to make suitable modifications. The current SRL profile somehow shows those kinds of SRL pattern information; however, it does not provide learners with insight or guidelines to make adjustments.

This study and its limitations suggest the following two future works. First is to consider online learning resources with the intentions, such as the purpose of use, timeframe, and lecture content that instructors and course designers implant in the resources. Second is about validation of the model. We plan to model students' learning history, then present the model to the students for their feedback and evaluation.

ACKNOWLEDGEMENT

This work was supported by Doctoral Research Fellowship (DRF) of JAIST.

REFERENCES

Bernacki, M. L., Aguilar, A. C., & Byrnes, J. P. (2011). Chapter 1. Self-regulated learning and technology-enhanced learning environments: An opportunity-propensity analysis. In G. Dettori & D. Persico (Eds.), *Fostering Self-Regulated Learning through ICT* (pp. 1–26). IGI Global. https://doi.org/10.4018/978-1-61692-901-5

Bertsekas, D. P., & Tsitsiklis, J. N. (2000). Markov chains. In *Introduction to Probability* (pp. 2–3). Massachusetts Institute of Technology, Cambridge, Massachusetts.

Geigle, C., & Zhai, C. (2017). Modeling MOOC Student Behavior With Two-Layer Hidden Markov Models. *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*, 205–208. https://doi.org/10.1145/3051457.3053986

Gilbert Strang. (2016). Chapter 6. Eigenvalues and Eigenvectors. In *Introduction to Linear Algebra* (5th ed., pp. 288–303). Wellesley-Cambridge Press.

Hao, J., Gan, J., & Zhu, L. (2022). MOOC performance prediction and personal performance improvement via Bayesian network. *Education and Information Technologies*. https://doi.org/10.1007/s10639-022-10926-8

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics dataset. *Scientific Data*, *4*(1). https://doi.org/10.1038/sdata.2017.171

Qiao, L., Zhao, W., & Xu, X. (2021). Mining and Analysis of Self-regulated Learning Process Model: Based on Hidden Markov Model. *2021 Tenth International Conference of Educational Innovation through Technology (EITT)*, 276–281. https://doi.org/10.1109/EITT53287.2021.00061

Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., Jiang, B., & Chen, P. (2022). Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports*, *12*(1), 453. https://doi.org/10.1038/s41598-021-03867-8

Spedicato, G. A. (2017). Discrete Time Markov Chains with R. *The R Journal*. https://journal.r-project.org/archive/2017/RJ-2017-036/index.html

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*. https://doi.org/10.1016/j.chb.2019.106189